

Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials

J. K. Mansour¹ · J. L. Beaudry² · R. C. L. Lindsay³

© The Author(s) 2017. This article is an open access publication

Abstract Eyewitness identification experiments typically involve a single trial: A participant views an event and subsequently makes a lineup decision. As compared to this single-trial paradigm, multiple-trial designs are more efficient, but significantly reduce ecological validity and may affect the strategies that participants use to make lineup decisions. We examined the effects of a number of forensically relevant variables (i.e., memory strength, type of disguise, degree of disguise, and lineup type) on eyewitness accuracy, choosing, and confidence across 12 target-present and 12 target-absent lineup trials ($N = 349$; 8,376 lineup decisions). The rates of correct rejections and choosing (across both target-present and target-absent lineups) did not vary across the 24 trials, as reflected by main effects or interactions with trial number. Trial number had a significant but trivial quadratic effect on correct identifications ($OR = 0.99$) and interacted significantly, but again trivially, with disguise type ($OR = 1.00$). Trial number did not significantly influence

participants' confidence in correct identifications, confidence in correct rejections, or confidence in target-absent selections. Thus, multiple-trial designs appear to have minimal effects on eyewitness accuracy, choosing, and confidence. Researchers should thus consider using multiple-trial designs for conducting eyewitness identification experiments.

Keywords Eyewitness identification · Simultaneous and sequential lineups · Multiple trials · Multilevel modelling · Eyewitness confidence

Eyewitness identification research has contributed significantly to judicial and policing practices over the past few decades and this momentum continues (Innocence Project, [n.d.](#); *State v. Henderson*, 2011; Technical Working Group for Eyewitness Evidence, 1999; National Research Council, 2014). A major challenge to eyewitness researchers is the balance between ecological validity and methodological rigor: Researchers must decide which aspects of their experimental design are critical for internal validity and which ought to closely resemble those encountered by real-world eyewitnesses to ensure external validity. In the real world, eyewitnesses typically see one crime and (may) participate in an identification procedure for a suspect. Translated into an experimental design, a participant is assigned to one experimental condition and views one mock crime and an accompanying lineup. This between-subjects design has the advantage of closely mirroring the real eyewitness experience but the disadvantage of producing only one recognition and one confidence data point per participant (cf. Brewer, Weber, Clark, & Wells, 2008).

Researchers obtaining a single data point per participant (per measure) require a large number of participants to obtain sufficient power to detect reliable differences. Concerns about power are heightened when the measure of interest is

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0855-0) contains supplementary material, which is available to authorized users.

✉ J. K. Mansour
jmansour@qmu.ac.uk

¹ Memory Research Group, Centre for Applied Social Sciences, Psychology & Sociology, Queen Margaret University, Edinburgh EH21 6UU, UK

² Centre for Forensic Behavioural Sciences, Department of Psychological Sciences, School of Health Sciences, Faculty of Health, Arts and Design, Swinburne University of Technology, Hawthorn VIC 3122, Victoria, Australia

³ Department of Psychology, Queen's University, Kingston K7L 3N6, Ontario, Canada

dichotomous (Tabachnick & Fidell, 2007), as is often the case in eyewitness research. Obviously, collecting data from larger samples requires more time and resources. Using a repeated-measures design to obtain multiple data points from each participant allows researchers to obtain greater power with smaller sample sizes. In addition, within-subjects designs in general have higher internal validity because each participant acts as their own control group (Glass & Hopkins, 1996). Thus, eyewitness researchers must decide between maximizing external validity by collecting one data point per participant from a large number of participants and maximizing power, internal validity, and resources by collecting multiple data points per participant. We considered how such repeated-measures designs affect eyewitness accuracy, choosing, and confidence.

Potential effects of multiple trials

Researchers may be suspicious of using multiple lineup trials because such a procedure may change how participant-eyewitnesses approach the lineup task and/or may produce practice effects (VanLehn, 1996). Participants in multiple-trial eyewitness identification experiments may become more or less accurate as they complete more trials because they become more aware of the task demands, develop beliefs about the researchers' hypotheses (Rosenthal, 1966), or adopt more or less stringent selection (or rejection) criteria.

Researchers interested in increasing children's identification accuracy have explored presenting practice trials before the "real" lineup. When presented with traditional identification procedures, children are less likely to reject lineups than adults—leading to a higher probability of false identifications when the perpetrator is absent (Fitzgerald & Price, 2015; Pozzulo & Lindsay, 1998). Pozzulo and Lindsay (1997) hypothesized that practice would decrease choosing by demonstrating to children that saying "no" to a lineup can be an appropriate response. Contrary to this intuitive hypothesis, explicit practice does not tend to improve children's accuracy (e.g., Goodman, Bottoms, Schwartz-Kenney, & Rudy, 1991; Parker & Myers, 2001; Parker & Ryan, 1993; Pozzulo & Lindsay, 1997). On the basis of their meta-analysis, Pozzulo and Lindsay (1998) concluded that practice likely has no effect on children's correct rejections although it may increase their correct identifications. If multiple-trial experiments produce higher correct identification rates, consumers of this research (police, lawyers, judges, researchers, etc.) may overestimate the accuracy of real-world child witnesses because they are unaware that this level of performance is an artifact of the research paradigm.

Research examining practice effects with adults is inconsistent. Shapiro and Penrod's (1986) meta-analysis of facial recognition and eyewitness identification studies found no consistent positive or negative effects of practice—which they referred to as training—on facial recognition and lineup decisions. Practice

reduced accuracy in some studies (e.g., Platz & Hosch, 1988), but had no effect in others (e.g., Malpass, Laviqueur, & Weldon, 1973). Although only a few studies have explicitly examined the issue of practice (eight studies with target-present lineups; five with target-absent), Shapiro and Penrod concluded that extensive practice (e.g., describing, recognizing, or comparing 90+ faces) is ineffective, but that short (20-minute) training programs may increase facial recognition accuracy. This pattern suggests that the "practice" inherent in a multiple-trial design may influence performance on early trials. However, one eyewitness study using multiple trials reported no learning effects across eight trials (Mansour, Lindsay, Brewer, & Munhall, 2009). Given the literature, we expected correct identifications to remain stable across multiple trials or, at most, to slightly increase in early trials and level off in later trials (i.e., a quadratic effect of trial number).

Possible interactions with multiple trials

Of particular concern with the use of multiple-trial experiments is whether the manipulated variables would interact with multiple trials to produce systematic changes in accuracy, choosing, or confidence. One such variable is lineup type. Sequential lineups involve presenting one lineup member at a time, combined with other procedural components designed to reduce false identifications relative to presenting lineup members simultaneously (simultaneous lineups; Lindsay, Mansour, Beaudry, Leach, & Bertrand, 2009). Normally, sequential lineups are backloaded, which implies to the eyewitness that the lineup contains more people than it actually does, in order to reduce pressure to choose someone (Horry, Palmer, & Brewer, 2012; Lindsay & Wells, 1985). In contrast, the number of lineup members in a simultaneous lineup is immediately obvious. Thus, eyewitnesses viewing simultaneous lineups may maintain a constant decision criterion (Meissner, Tredoux, Parker, & MacLin, 2005), whereas eyewitnesses viewing sequential lineups may adopt a more lenient criterion as they become familiar with the size of the lineups. If this is the case, a multiple-trial approach would be inadvisable with sequential lineups.

Prior research examining how knowledge of the nominal size of the sequential lineup affects identification decisions speaks to this possibility. Lindsay, Lea, and Fulford (1991) and Horry et al. (2012) found correct rejections (but not correct identifications) were lower when participants were aware of how many members comprised their sequential lineup. A participant who rejected a sequential lineup in a multiple-trial experiment would discover how many lineup members comprised the lineups, which could lead them to adopt a more lenient criterion for identification as they approached the end of subsequent lineups. If selecting a lineup member terminated the trial (as in the present research), selections would not provide information about lineup size. Thus, participants who

view sequential lineups may make more correct rejections in early trials than in later trials, when size may become apparent and, thus, perceived pressure to choose increases. We expected no significant difference in correct rejections across trials with simultaneous lineups given the stability of apparent lineup size with these lineups.

A second factor that could interact with trial number is the strength of the eyewitness' memory for the perpetrator. The difficulty of a lineup decision is at least partially a function of memory trace strength. People with a weak memory trace—due to poor encoding conditions or delay—tend to perform worse when they view multiple lineups for the same suspect (Godfrey & Clark, 2010; Lindsay, Mansour, Kalmet, Bertrand, & Melsom, 2011). Palmer, Brewer, and Weber (2010) suggested that viewing multiple, non-independent lineups for the same target negatively affects metacognitions about memory strength. When they led participants to believe their memory strength was poor (on the basis of feedback or presentation of a second lineup), performance on the second lineup suffered (fewer correct identifications and correct rejections). Palmer et al.'s conclusion is consistent with research showing that people use perceived prior success or failure to inform judgments of future success (Feather, 1966). Thus, a participant's experience with prior lineups and/or their perceived memory strength may inform subsequent lineup decisions.

As a consequence for multiple-trial designs, an important consideration is whether confidence interacts with willingness to choose from lineups. When one's memory for a face is weak, confidence should be low. In contrast, when one's memory is strong, selections are likely to be made with confidence. A feeling of confidence in a particular decision (arising from the decision being "easy") may lead participants to have more confidence in their *ability* to remember, potentially leading them to become more willing to identify someone from the lineup. That is, previous (apparent) success may increase perceptions of one's general ability to make correct memory-based decisions. Generally, confidence in identification decisions has been associated with conditions that influence memory strength: the relation between confidence and accuracy deteriorates as viewing conditions deteriorate (Lindsay, Read, & Sharma, 1998). Using a multiple-trial design in which some conditions make it difficult to form a good memory trace (e.g., disguise) may eliminate systematic changes in confidence in identifications.

Given that the relationship between confidence and accuracy is weak or nonexistent for lineup rejections (Leippe & Eisenstadt, 2007) we would logically expect no systematic effects of these variables on confidence in rejections. Accordingly, a repeated-measures design that provides a randomized mix of target-present and target-absent lineup trials may not lead participants to become

more (or less) confident in their ability; thus, choosing, accuracy, and confidence may be unaffected by trial.

Present study

Whether multiple eyewitness identification trials for independent targets influence the validity of conclusions is an empirical question that has not been addressed directly. We examined whether correct identifications, correct rejections, choosing, and decision confidence changed over 24 trials, and examined possible interactions with lineup type and memory strength variations. The [supplemental materials](#) also present analyses of overall accuracy—defined as the proportion of all lineup decisions that were correct (i.e., correct identifications and correct rejections)—and of mean overall confidence. Critically, an absence of effects of trial number and interactions of trial number with other variables (e.g., memory strength) can be taken as evidence that multiple-trial experiments do not compromise the validity of identification data.

Method

We reanalyzed data reported in Mansour et al. (2012) and included newly collected data using an almost identical methodology (total $N = 8,376$ lineup decisions). We summarize only the key factors of the earlier work; readers are encouraged to refer to Mansour et al. for a more detailed methodology. Participants completed 24 trials in which they watched a mock-crime video, made a lineup decision after no delay or a short delay (see below), and reported their confidence in that decision on a scale from 0% (not at all confident) to 100% (extremely confident). All participants received fair lineup instructions (Malpass & Devine, 1981) in conjunction with each lineup.

With the exception of memory strength (detailed later), all data sets included the same manipulations so we report them together in this section. Participants were randomly assigned to all between-subjects manipulations and counterbalancing was employed for within-subjects manipulations. First, we manipulated between subjects *lineup type*. Lineup type refers to whether the lineups viewed by participants were simultaneous (i.e., all lineup members presented at once) or sequential (i.e., lineup members presented one at a time, a response for each lineup member required before viewing the next lineup member, no indication given as to the total number of lineup members; Lindsay et al., 2009). Second, we manipulated between subjects *type of disguise* (toque and sunglasses versus stocking mask). Disguised targets wore a toque (i.e., knitted hat or beanie) and/or sunglasses or they wore a stocking mask covering variable portions of

their face. Third, we manipulated within subjects *degree of disguise*. Toque and sunglasses participants viewed targets that were undisguised, wore a toque, wore sunglasses, or wore both a toque and sunglasses. Stocking participants viewed targets who were undisguised, wore a stocking covering their hair and forehead (1/3 covered), wore a stocking covering to just below their nose (2/3 covered), or wore a stocking completely covering their head (fully covered). Fourth, we manipulated within subjects *target presence*—that is, whether a particular lineup contained the target (target-present) or not (target-absent).

The mock-crime videos were designed to elicit different levels of memory trace strength for the target. The mock-crime videos intended to elicit good memory strength were approximately 30 s in length and filled a 19-in. monitor. To produce mock-crime videos that would elicit a moderate strength memory, we shortened the 30-s videos to 3 s and resized them to fill one-third of a 19-in. monitor. To produce a poor strength memory, we modified the moderate memory strength condition by including a 30-s delay in which participants completed a visual search task between viewing each 3-s mock-crime video and its associated lineup. We added this delay to further weaken participants' memory for each target relative to the other conditions by allowing an opportunity for forgetting to occur. With the exception of these changes, the mock-crime videos across memory strength conditions were identical. Our correct identification rates support our good, moderate, and poor memory strength categorization (see the Results section).

Each mock-crime video depicted one of four mock crimes: discussion of a bank robbery, a plot to murder someone, planning of a burglary with an off-screen accomplice, or questioning by an off-screen police officer after a robbery. All videos displayed one target from the shoulders up and the targets followed the same script for the respective mock crimes. Thus, the mock-crime videos depicting the same mock crime were identical except for the target. The videos chosen were selected from a larger pool of videos (approximately 35) based on ease of producing a lineup for the target (e.g., we opted not to include targets for which the pool of filler photographs was small) and to ensure equal numbers of male and female targets.

All lineups included six facial pictures (neck up and thus no clothing cues) of people looking straight into the camera. Fillers were selected by using an iterative match to description procedure (Turtle, Lindsay, & Wells, 2003) within the limits of the pictures available within the lab. Five members of the target-absent lineups were used as fillers in the target-present lineups. No photo appeared in any lineup for more than one target. All lineup members were undisguised.

Mansour et al. (2012) data set

The participants in the two experiments from Mansour et al. (2012) were students at an Eastern Canadian University. The participants in Experiment 1 ($N = 98$) were randomly assigned to the toque and/or sunglasses condition of our disguise type manipulation, whereas the participants in Experiment 2 ($N = 102$) were randomly assigned to the stocking condition of our disguise type manipulation.¹ In addition to the manipulations described above, participants were expected to have different levels of memory strength, as described above.

Good memory strength Approximately two-thirds of the participants in Experiment 1 ($n = 56$) and Experiment 2 ($n = 58$) participated in the good memory strength condition. The quality of their exposure to the target presumably resulted in a good opportunity to encode his or her face. Lineups were presented immediately after the videos, providing little to no opportunity for forgetting.

Moderate memory strength The other participants in Experiment 1 ($n = 38$) and Experiment 2 ($n = 39$) participated in the moderate memory strength condition. This exposure presumably resulted in a moderate opportunity to encode the perpetrator, with little or no opportunity for forgetting.

Additional data set (poor memory strength)

We later collected additional data from 158 participants (randomly assigned to the toque and/or sunglasses disguise [$n = 78$] or stocking disguise [$n = 80$] conditions) at a Western Canadian university using nearly the same procedures as Mansour et al. (2012).

Participants In this additional dataset, the participants were predominantly female (.68) and categorized themselves as Asian (.52), White (.25), or other (.23). Most participants were of college age ($M = 20.36$, $SD = 2.75$, range = 17–38). All participants received introductory psychology course credit in exchange for participating.

Design Approximately equal proportions of participants were randomly assigned to the between-subject manipulations of lineup type (simultaneous, sequential) and disguise type (toque/sunglasses, stocking). As in Mansour et al. (2012),

¹ The data from four participants from Experiment 1 and five participants from Experiment 2 of Mansour et al. (2012) were not included in these analyses because they completed 32 (rather than 24) trials. As such, these participants saw different targets for the first 24 trials than were seen by the other participants. We excluded them from these analyses to remove additional noise from the data. The following *ns* reported represent those used in our analyses.

we manipulated degree of disguise and target presence within subjects and employed counterbalancing for these variables.

Materials and procedure The materials (including videos and lineups) and procedure were the same as in Mansour et al. (2012), with the exception of the 30-s delay between videos and lineups, during which participants viewed a Where's Waldo² image and answered related questions (e.g., "How many people are sunburned in this picture?").

Measures

We focused our analyses on six measures (three identification decisions, three confidence decisions); additional analyses are available in the [supplemental materials](#). For target-present lineups, we coded target identifications as accurate (correct identifications) and coded responses of "not there" and selections of lineup fillers as inaccurate (inaccurate target-present decisions). When reporting descriptive statistics alongside our model results, we provide the proportion of correct identifications (number of correct identifications divided by the total number of target-present trials). For target-absent lineups, we avoided the issue of designating an innocent suspect because the discipline lacks a consistent method. We coded "not there" responses as accurate (correct rejections) and all selections as inaccurate (target-absent selections). The relevant descriptive statistic for this measure is the proportion of correct rejections calculated as the number of correct rejections divided by the total number of target-absent trials. Overall accuracy was also calculated (see the [supplemental materials](#)) as the number of correct responses (i.e., correct identifications plus correct rejections) divided by the total number of trials. Identification decisions were also coded as selections (correct identifications, filler selections, and target-absent selections) or rejections (incorrect rejections and correct rejections) so we could analyze choosing. We report the proportion of choices with our inferential results (i.e., the number of selections divided by the total number of trials). Finally, we examined confidence in correct identifications, confidence in correct rejections, and confidence in target-absent selections. The [supplemental materials](#) report analyses of overall confidence, defined as mean confidence across all lineup decisions. Intervals reported after proportions and means are 95% confidence intervals.

Analytic approach

Each participant made multiple lineup decisions and we did not randomly assign participants to the good, moderate, or

poor memory strength conditions. As such, lineup decisions were nested within participants, which were nested within memory strength. Thus, our data were nested in three levels with trials at Level 1, participants at Level 2, and memory strength at Level 3. Participants were randomly assigned to the between-subject conditions of lineup type and disguise type, while degree of disguise was manipulated within subjects; therefore, these manipulations were incorporated at Level 1 (Field, 2009).

We used multilevel mixed-effects modeling to evaluate models for the six measures described above. First we modeled how participants responded on each trial: one set of models aimed to predict correct identifications, one set aimed to predict correct rejections, and one set aimed to predict choosing. The remaining models aimed to predict participants' confidence on a particular trial given their specific decision. That is, we modeled confidence in correct identifications, confidence in correct rejections, and confidence in target-absent selections.

[Appendix A](#) provides the mathematical formulas for each model and [Appendix B](#) provides a flowchart of the process we engaged in for modeling, but we outline the general logic here. We first determined whether a multilevel model was necessary by comparing a one-level model with no predictors to two- and then three-level models with no predictors (i.e., different null models). The three-level model was normally superior to a one- or two-level model except in two cases (noted in the Results); therefore, we do not discuss the comparisons of different null models in the results.

Next, and in keeping with standard modeling procedures, we examined which fixed effects should be included in the multilevel model (Field, 2009). Each fixed effect was added to the null model (Model 1) individually, and the fit was compared to the null model (Models 2–5). Thus, we compared the null model to four separate fixed-effect models; each model included one of lineup type, disguise type, degree of disguise, or trial number. Fixed effects resulting in superior model fit were next included together in a model to test whether a model with multiple predictors (Model 6) improved fit relative to the null or to models with each individual predictor. If one or none of the fixed effects improved the fit, then we could not construct a Model 6, and so proceeded to the next step. Once the best-fitting fixed-effects model was determined, it was compared to models including interactions of trial number with the fixed effects (Models 7–9) and with the nesting variables (memory strength and participant; Models 10 and 11, respectively). For models with interactions, relevant fixed effects were always included.

We interpreted our results using the likelihood ratio test because this is the most liberal test and the most widely reported one in the eyewitness field (e.g., Horry, Halford, Brewer, Milne, & Bull, 2014; Wright & London, 2009).

² TM & © 2008 Entertainment Rights Distribution Limited. All rights reserved.

This test compares the log-likelihood of nested models on a chi-square distribution (Hox, 2010). In addition, we report the Akaike information criterion (AIC), Akaike weights (w_i), and Bayes factor (BF) for interested readers. Wagenmakers and Farrell (2004) recommend converting raw AIC values to Akaike weights to obtain an approximate measure of the probability that the model at hand is the best of the various models considered. Akaike weights vary between 0 and 1; therefore, an Akaike weight of .56 indicates that the specific model has a 56% chance of being the best model out of the set of models considered for predicting the variance in a given dataset. Within the text of the results, we calculated a ratio of the Akaike weights for the models being compared. The more complex model was always in the numerator and the simpler model (null or the best-fitting to that point) in the denominator; therefore, ratios greater than 1 indicate evidence for the more complex model. Finally, the BF, the most conservative test, was estimated from the Bayesian information criterion (BIC; Jarosz & Wiley, 2014). Critically for our interest in potential null effects, the BF allows one to test for evidence of null results (Jarosz & Wiley, 2014). For all BF calculations, as we did with the Akaike weight ratios, we placed the more complex model in the numerator and the simpler model in the denominator. Thus, a BF greater than 1 provides evidence in favor of the more complex model. When the BIC is used to calculate the BF, the model is more likely to favor the null hypothesis (i.e., the simpler model, in this case) over the alternative hypothesis (i.e., the more complex model, in this case); thus, models should be evaluated on the basis of both AIC and BF (Weakliem, 1999).

Results

Correct identifications

We created a model to examine which of our experimental manipulations best predicted the accuracy of decisions from target-present lineups (correct identifications versus inaccurate target-present decisions). This analysis, however, revealed that the data were not best modeled using the nested analysis. In fact, when examining models without predictors, the model with an intercept at Level 1 only was better than the models with random intercepts at Levels 1 and 2, or at Levels 1, 2, and 3. Thus, we used hierarchical logistic regression to analyze these data.

In the first step, we entered all of the fixed effects: memory strength, lineup type, disguise type, degree of disguise, and trial number. In Step 2, we entered the interactions of interest: the interactions of trial number with each of memory strength, lineup type, disguise type, and degree of disguise. The overall model was significant, $\chi^2(9) = 136.57$, $p < .001$, $R_{\text{Nagelkerke}}^2 = .04$; with Step 1 significant, $\chi^2(5) = 130.10$, $p < .001$, $R_{\text{Nagelkerke}}^2 = .04$; but not Step 2, $\chi^2(4) = 6.47$,

$p = .17$, $R_{\text{Nagelkerke}}^2 = .002$ (none of the predictors in Step 2 were significant, with all $ps > .08$).

Memory strength, lineup type, and disguise type were significant predictors in Step 1. Participants made more correct identifications in the good memory strength condition ($M = .72$ [.70, .74]) than in the moderate ($M = .65$ [.62, .68]) and poor ($M = .58$, [.55, .60]) memory strength conditions, $B = .34$, $\chi^2(1) = 79.32$, $p < .001$, $OR = 1.40$. All three levels differed significantly from each other using a Bonferroni correction ($ps \leq .001$). Simultaneous lineups ($M = .68$ [.66, .70]) led to more correct identifications than did sequential lineups ($M = .59$ [.57, .62]), $B = .42$, $\chi^2(1) = 41.56$, $p < .001$, $OR = 1.53$. Finally, participants made more correct identifications in the toque and/or sunglasses conditions ($M = .66$ [.64, .68]) than in the stocking conditions ($M = .62$ [.60, .64]), $B = .19$, $\chi^2(1) = 8.63$, $p < .001$, $OR = 1.21$. Neither degree of disguise, $B = .05$, $\chi^2(1) = 3.02$, $p = .08$, $OR = 1.05$, nor trial number, $B = -.003$, $\chi^2(1) = 0.45$, $p = .50$, $OR = 1.00$ (see Fig. 1A), was a significant predictor. Thus, the hierarchical logistic regression revealed that memory strength, lineup type, and disguise type influenced correct identifications, but that degree of disguise and trial number did not.

Because we anticipated there might be an increase in correct identifications early on that would stabilize in later trials, we mean-centered and then squared trial number to create the quadratic effect of trial number. The overall model was significant, $\chi^2(10) = 201.02$, $p < .001$, $R_{\text{Nagelkerke}}^2 = .06$, with Step 1 being significant, $\chi^2(6) = 193.74$, $p < .001$, $R_{\text{Nagelkerke}}^2 = .06$, but not Step 2, $\chi^2(4) = 7.28$, $p = .12$, $R_{\text{Nagelkerke}}^2 = .002$. The conclusions were similar for the model with the linear effect of trial number with two exceptions. In Step 1, the quadratic effect of trial number, $B = -.006$, $\chi^2(1) = 63.41$, $p < .001$, $OR = 0.99$, was significant (see Fig. 1B). Second, although Step 2 did not account for significant variance, the interaction of the quadratic effect of trial number with disguise type was a significant predictor in Step 2, $B = -.004$, $\chi^2(1) = 6.67$, $p = .01$, $OR = 1.00$ (see Fig. 1B).

Finally, we ran a multilevel model to examine the effects of trial number (linear and quadratic) on correct identifications. Correct identifications in a three-level nested model were best modeled by including lineup type, disguise type, the quadratic effect of trial number, and the interaction of the quadratic effect of trial number and disguise type (see Table 1), consistent with the results of the logistic regression. Thus, the results indicate a quadratic relationship between trial number and correct identifications, although the effect size is small. We consider the implications of these trial effects in the Discussion section.

Correct rejections

In this analysis, we predicted the accuracy of lineup decisions from target-absent lineups (correct rejections versus target-

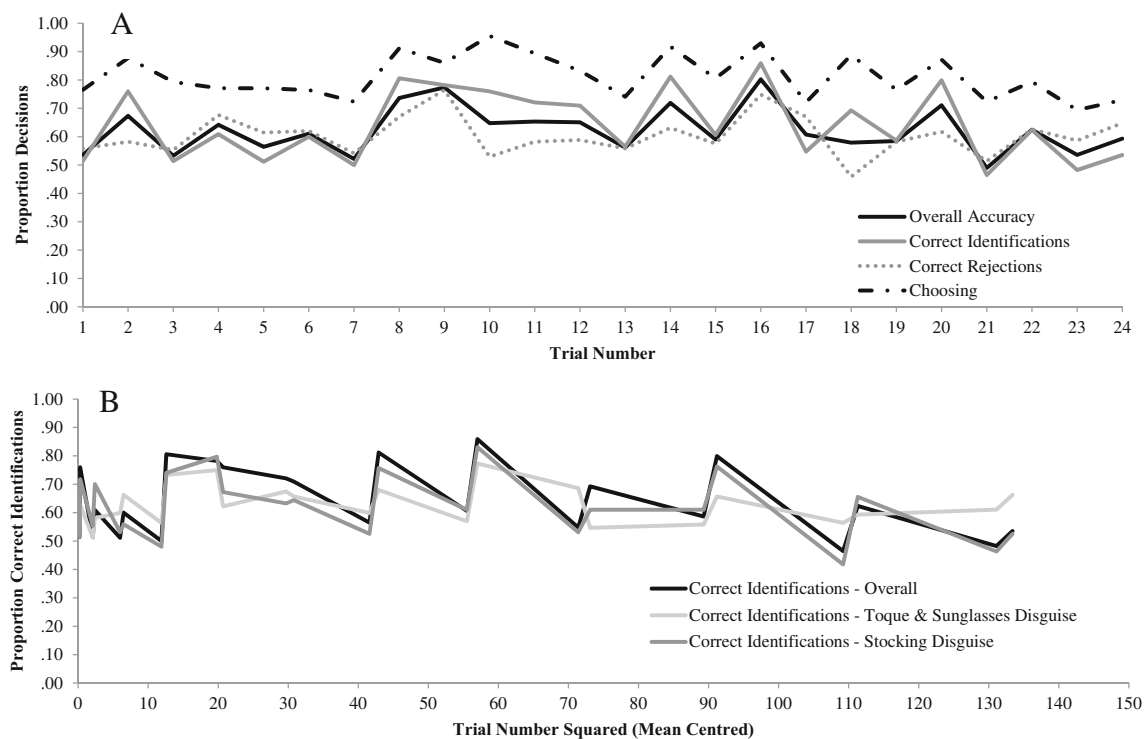


Fig. 1 (A) Proportions of actual overall accuracy (i.e., proportion of lineup decisions resulting in either correct identifications or correct rejections), correct identifications, correct rejections, and choosing, by trial number, and (B) proportion of correct identifications by trial number squared (i.e., a quadratic effect)

absent selections). Accuracy was highest when memory strength was good ($M = .74$ [.72, .76]), followed by moderate ($M = .66$ [.63, .69]), and poor ($M = .48$ [.46, .50]).

Including lineup type improved the fit relative to the null model (see Table 2), $\chi^2(1) = 20.20$, $p < .001$, w_i ratio = 8,955.29, $BF = 365.04$, such that more correct rejections were made when participants viewed sequential ($M = .66$ [.64, .68]) than when they viewed simultaneous lineups ($M = .55$ [.53, .57]). The fit relative to the null model was not improved by including disguise type, $\chi^2(1) = 1.40$, $p = .24$, w_i ratio = 0.74, $BF = 0.03$; degree of disguise, $\chi^2(1) = 0.20$, $p = .65$, w_i ratio = 0.39, $BF = 0.02$; or trial number, $\chi^2(1) = 0.20$, $p = .65$, w_i ratio = 0.41, $BF = 3.11 \times 10^{-14}$ (see Fig. 1A) on their own.

Next we considered whether including interactions with trial number (separately) improved fit relative to the model with lineup type only. Neither the interaction of trial number and lineup type, $\chi^2(2) = 0.40$, $p = .82$, w_i ratio = 0.17, $BF = 2.89 \times 10^{-4}$; nor the interaction of trial number and disguise type, $\chi^2(3) = 1.80$, $p = .61$, w_i ratio = 0.12, $BF = 8.72 \times 10^{-6}$; nor the interaction of trial number and degree of disguise, $\chi^2(3) = 0.60$, $p = .90$, w_i ratio = 0.07, $BF = 5.29 \times 10^{-6}$, improved model fit.

Allowing the slopes of correct rejections across trial numbers to vary across memory strength conditions did not improve model fit, $\chi^2(4) = 0.20$, $p = 1.00$, w_i ratio = 0.02, $BF = 6.49 \times 10^{-8}$. Allowing each participant to have a different

slope of correct rejections across trial numbers also did not improve the model fit, $\chi^2(5) = 1.20$, $p = .94$, w_i ratio = 0.01, $BF = 1.61 \times 10^{-9}$, and this model did not converge. Thus, the best-fitting model for correct rejections was a three-level model with the fixed effect of lineup type only. Neither trial number nor any interactions with trial number predicted participants' correct rejections.

Choosing

In this analysis, we considered whether the manipulated or nested variables predicted participants' selections from the lineup, correctly or incorrectly, regardless of whether or not the lineup contained the target. Table 3 depicts the model parameter estimates and fit indices. A three-level model was preferable to a two- or one-level model. Participants were less likely to choose someone from the lineup when their memory strength was good ($M = .54$ [.52, .56]), followed by moderate ($M = .56$ [.54, .59]), and poor ($M = .67$ [.66, .69]).

In comparison to the null model, separately adding lineup type, $\chi^2(1) = 23.80$, $p < .001$, w_i ratio = 5.42×10^4 , $BF = 1.64 \times 10^3$, or disguise type, $\chi^2(1) = 5.80$, $p = .02$, w_i ratio = 6.36, $BF = 0.19$, significantly improved the fit, whereas separately adding degree of disguise, $\chi^2(1) = 1.20$, $p = .27$, w_i ratio = 0.67, $BF = 0.02$, or trial number, $\chi^2(1) = 2.00$, $p = .16$, w_i ratio = 0.95, $BF = 0.03$ (see Fig. 1A), had no effect. The model with both disguise type and lineup type significantly improved

Table 1 Parameter estimates for predictors in models of correct identifications (4,188 observations)

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8 [†]	Model 9	Model 10 [‡]	Model 11 [‡]
Fixed Effects											
Intercept	0.65 (0.03)	0.70 (0.04)	0.67 (0.04)	0.62 (0.04)	0.72 (0.04)	0.79 (0.04)	0.79 (0.04)	0.76 (0.04)	0.76 (0.05)	0.76 (0.30)	0.76 (0.07)
Lineup type	–0.10 (0.02)	–0.10 (0.02)	–0.04 (0.02)			–0.09 (0.02)	–0.10 (0.02)	–0.09 (0.02)	–0.09 (0.02)	–0.10 (0.02)	–0.10 (0.02)
Disguise type						–0.04 (0.02)	–0.04 (0.02)	0.003 (0.02)	–0.04 (0.02)	0.003 (0.02)	0.003 (0.02)
Degree of disguise											
Trial number				0.01 (0.01)	<0.001 (0.001)	<0.001 (0.001)	<0.001 (0.001)	<0.001 (0.001)	0.01 (0.01)	<0.001 (0.001)	<0.001 (0.001)
Trial number squared					–0.001 (<0.001)	–0.001 (<0.001)	–0.001 (<0.001)	–0.001 (<0.001)	–0.001 (<0.001)	–0.001 (<0.001)	<0.001 (<0.001)
Lineup Type × Trial Number squared							<0.001 (<0.001)				
Disguise Type × Trial Number squared								–0.001 (<0.001)		–0.001 (<0.001)	
Degree of Disguise × Trial Number squared											
Random Parameters											
Level 2 intercept variance (participant)	0.01 (0.11)	0.01 (0.10)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)	0.01 (0.10)	0.01 (0.08)
Level 2 slope variance (participant)											<0.001 (<0.001)
Level 3 intercept variance (memory strength)	0.003 (0.06)	0.004 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.004 (0.06)	0.004 (0.06)	0.004 (0.06)	0.004 (0.06)	0.17 (0.41)	0.01 (0.12)
Level 3 intercept variance (memory strength)*											
Level 3 slope variance (memory strength)										0.10 (0.32)	
Model Fit											
Model df	4	5	5	5	6	8	9	9	10	12	13
Test change in df	–	1 ^a	1 ^a	1 ^a	2 ^a	4 ^a 3 ^{bc} 2 ^{de} 1 ^{fg}	1 ^h	1 ^h	2 ^h	3 ⁱ	4 ⁱ
Bayes Factor	–	1.63 × 10 ^{4a}	0.30 ^a	0.07 ^a	9.72 × 10 ^{10a}	4.32 × 10 ^{14a} 2.65 × 10 ^{10b} 1.44 × 10 ^{15c} 4.447.07 ^d 9.72 × 10 ^{10e} 1.13 × 10 ^{11f} 1.48 × 10 ^{8g}	0.02 ^h	0.95 ^h	0.001 ^h	2.64 × 10 ^{–8i}	5.59 × 10 ^{–8i}
AIC	5,630.1	5,604.3	5,626.1	5,628.9	5,566.8	5,537.2	5,539.2	5,531.1	5,538.0	5,546.9	5,539.1
Akaike weights	2.58 × 10 ^{–22}	1.14 × 10 ^{–16}	2.11 × 10 ^{–21}	5.20 × 10 ^{–22}	1.59 × 10 ^{–8}	.04	.02	.90	.03	3.32 × 10 ^{–4}	.02
–2 log likelihood	5,622.1	5,594.3	5,616.1	5,618.9	5,554.8	5,521.2	5,521.2	5,513.1	5,518.0	5,522.9	5,513.1

Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] Best-fitting model. [‡] Models that did not converge; therefore, the values should be interpreted with caution. Superscripts indicate df and Bayes factor for comparison between the current model and ^aModel 1 (null model), ^bModel 2, ^cModel 3, ^dModel 5, ^ea model comprising the same parameters as Model 6 without trial number and trial number squared, ^fa model comprising the same parameters as Model 6 without disguise type, ^ga model comprising the same parameters as Model 6 without lineup type, ^hModel 6, and ⁱModel 8. *Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in [Appendix A](#).

Table 2 Parameter estimates for predictors in models of correct rejections (4,188 observations)

Predictor	Model 1	Model 2 [†]	Model 3	Model 4	Model 5	Model 7	Model 8	Model 9	Model 10	Model 11 [‡]
Fixed Effects										
Intercept	0.63 (0.06)	0.58 (0.06)	0.61 (0.06)	0.62 (0.06)	0.63 (0.06)	0.59 (0.06)	0.56 (0.07)	0.56 (0.07)	0.58 (0.06)	0.58 (0.06)
Lineup type		0.10 (0.02)				0.09 (0.03)	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)	0.10 (0.02)
Disguise type			0.03 (0.02)				0.04 (0.03)			
Degree of disguise				0.002 (0.01)				0.01 (0.01)		
Trial number					<0.001 (0.001)	−0.001 (0.001)	<0.001 (0.001)	0.001 (0.002)	<0.001 (0.001)	<0.001 (0.001)
Lineup Type × Trial Number						0.001 (0.002)				
Disguise Type × Trial Number							−0.001 (0.002)			
Degree of Disguise × Trial Number								<0.001 (0.001)		
Random Parameters										
Level 2 intercept variance (participant)	0.03 (0.18)	0.03 (0.17)	0.03 (0.18)	0.03 (0.18)	0.03 (0.18)	0.03 (0.17)	0.03 (0.17)	0.03 (0.17)	0.03 (0.17)	0.04 (0.19)
Level 2 slope variance (participant)										<0.001 (0.004)
Level 3 intercept variance (memory strength)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.09)	0.01 (0.10)
Level 3 intercept variance (memory strength)*								0.003 (0.05)		
Level 3 slope variance (memory strength)								<0.001 (<0.001)	<0.001 (<0.001)	<0.001 (<0.001)
Model Fit										
Model df	4	5	5	5	5	7	8	8	9	10
Test change in df	–	1 ^a	1 ^a	1 ^a	1 ^a	2 ^b	3 ^b	3 ^b	4 ^b	5 ^b
Bayes Factor		365.04 ^a	0.03 ^a	0.02 ^a	3.11 × 10 ^{−14a}	2.89 × 10 ^{−4b}	8.72 × 10 ^{−6b}	5.29 × 10 ^{−6b}	6.49 × 10 ^{−8b}	1.61 × 10 ^{−9b}
AIC	5,405.8	5,387.6	5,406.4	5,407.7	5,407.6	5,391.2	5,391.9	5,392.9	5,395.3	5,396.4
Akaike weights	8.06 × 10 ^{−5}	.72	5.97 × 10 ^{−5}	3.12 × 10 ^{−5}	3.27 × 10 ^{−5}	.12	.08	.05	.02	.01
−2* log likelihood	5,397.8	5,377.6	5,396.4	5,397.7	5,397.6	5,377.2	5,375.9	5,376.9	5,377.3	5,376.4

Note: Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] indicates the best-fitting model.

[‡] Model failed to converge; therefore the values should be interpreted with caution. Superscripts indicate df and Bayes factor for the comparison between the current model and ^a Model 1 (null model) and ^b Model 2. Model 6 is not included, since there was no need to test a model with multiple fixed effects; that is, only one fixed effect improved the model fit relative to the null model. *Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in Appendix A.

Table 3 Parameter estimates for predictors in models of choosing (8,376 observations)

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 [†]	Model 7	Model 8	Model 9	Model 10	Model 11
Fixed Effects											
Intercept	0.59 (0.03)	0.63 (0.03)	0.61 (0.04)	0.60 (0.04)	0.60 (0.04)	0.65 (0.03)	0.66 (0.04)	0.67 (0.04)	0.70 (0.04)	0.66 (0.03)	0.66 (0.03)
Lineup type		−0.08 (0.02)				−0.08 (0.02)	−0.08 (0.02)	−0.08 (0.02)	−0.08 (0.02)	−0.08 (0.02)	−0.08 (0.02)
Disguise type			−0.04 (0.02)			−0.04 (0.02)	−0.04 (0.02)	−0.05 (0.02)	−0.04 (0.02)	−0.04 (0.02)	−0.04 (0.02)
Degree of disguise				−0.005 (0.004)					−0.01 (0.01)		
Trial number					−0.001 (0.001)		−0.001 (0.001)	−0.001 (0.001)	−0.003 (0.002)	−0.001 (0.001)	−0.001 (0.001)
Lineup Type × Trial Number							<0.001 (0.001)				
Disguise Type × Trial Number								0.001 (0.001)			
Degree of Disguise × Trial Number											
Random Parameters											
Level 2 intercept variance (participant)	0.02 (0.13)	0.02 (0.12)	0.02 (0.13)	0.02 (0.13)	0.02 (0.13)	0.01 (0.12)	0.01 (0.12)	0.01 (0.12)	0.01 (0.12)	0.01 (0.12)	0.02 (0.14)
Level 2 slope variance (participant)											<0.001 (0.005)
Level 3 intercept variance (memory strength)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.06)	0.003 (0.05)	0.003 (0.05)
Level 3 intercept variance (memory strength)*										<0.001 (<0.001)	<0.001 (<0.001)
Level 3 slope variance (memory strength)										<0.001 (<0.001)	<0.001 (<0.001)
Model Fit											
Model df	4	5	5	5	5	6	8	8	9	10	11
Test change in df	−	1 ^a	1 ^a	1 ^a	1 ^a	2 ^a	2 ^d	2 ^d	3 ^d	4 ^d	5 ^d
Bayes Factor	−	1635.98 ^a	0.19 ^a	0.02 ^a	0.03 ^a	284.29 ^a	3.35 × 10 ^{−4d}	3.53 × 10 ^{−4d}	1.06 × 10 ^{−5d}	3.94 × 10 ^{−8d}	5.56 × 10 ^{−6d}
AIC	11,416.9	11,395.1	11,413.2	11,417.7	11,417.0	11,391.6	11,393.6	11,393.5	11,393.4	11,397.6	11,396.6
Alkaike weight	7.18 × 10 ^{−7}	.04	4.57 × 10 ^{−6}	4.81 × 10 ^{−7}	6.83 × 10 ^{−7}	.22	.08	.09	.09	.01	.02
−2* log likelihood	11,408.9	11,385.1	11,403.2	11,407.7	11,407.0	11,379.6	11,377.6	11,377.5	11,375.4	11,377.6	11,374.6

Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] Best-fitting model. Superscripts indicate df and Bayes factor for the comparison between the current model and ^aModel 1 (null model), ^bModel 2, ^cModel 3, and ^dModel 6. *Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in [Appendix A](#).

fit relative to the models with only lineup type, $\chi^2(1) = 5.60$, $p = .02$, w_i ratio = 5.75, BF = 0.10, and only disguise type, $\chi^2(1) = 23.60$, $p < .001$, w_i ratio = 4.90×10^4 , BF = 544.57. Thus, the best-fitting fixed-effects model included disguise type and lineup type. Choosing was higher when participants viewed simultaneous ($M = .65$ [.63, .66]) rather than sequential ($M = .56$ [.54, .57]) lineups. Likewise, choosing was higher in the toque and/or sunglasses disguise condition ($M = .62$ [.61, .64]) than in the stocking disguise condition ($M = .58$ [.57, .60]).

We next examined whether separately including the three two-way interactions of trial number with lineup type, disguise type, and degree of disguise improved the fit. In sequence, we compared each of these predictors to the best-fitting fixed-effects model, containing lineup type and disguise type. None of the interactions improved the model fit. Specifically, relative to the best-fitting fixed-effects model, the interactions of trial number with lineup type, $\chi^2(2) = 2.00$, $p = .37$, w_i ratio = 0.37, BF = 3.35×10^{-4} ; disguise type, $\chi^2(2) = 2.20$, $p = .33$, w_i ratio = 0.39, BF = 3.53×10^{-4} ; and degree of disguise, $\chi^2(3) = 4.20$, $p = .24$, w_i ratio = 0.41, BF = 1.06×10^{-5} , did not improve the fit.

Our final step was to examine whether allowing the slope of trial number to vary over levels of memory strength or participants improved the fit. Neither modification to the best-fitting fixed-effects model improved the fit: $\chi^2(4) = 2.00$, $p = .74$, w_i ratio = 0.05, BF = 3.94×10^{-8} (memory strength), and $\chi^2(5) = 5.00$, $p = .42$, w_i ratio = 0.08, BF = 5.56×10^{-6} (participant). Thus, choosing was best predicted by a three-level model including the fixed effects of disguise type and lineup type. Neither trial number nor any interactions with trial number predicted participants' choosing.

Confidence in correct identifications

The three-level null model predicted confidence in correct identifications better than a one-level null model, but it did not differ from the two-level model. Since a three-level model most accurately reflects our data's structure, we continued with the three-level model as our null model. Confidence was highest for participants in the good memory strength condition ($M = 78.94\%$ [77.76, 80.12]), followed by the moderate ($M = 76.48\%$ [74.87, 78.09]), and poor conditions ($M = 75.45\%$ [74.03, 76.87]).

Lineup type, $\chi^2(1) = 12.20$, $p < .001$, w_i ratio = 164.02, BF = 8.58, but not disguise type, $\chi^2(1) = 1.20$, $p = .27$, w_i ratio = 0.61, BF = 0.03; degree of disguise, $\chi^2(1) = 1.20$, $p = .27$, w_i ratio = 0.67, BF = 0.04; or trial number, $\chi^2(1) = 1.60$, $p = .21$, w_i ratio = 0.74, BF = 0.04 (see Fig. 2), improved the model fit when entered separately and compared to the null model. Confidence in correct identifications was higher when participants viewed sequential ($M = 79.41\%$ [78.25, 80.57]) rather than simultaneous ($M = 75.02\%$ [73.91, 76.13]) lineups.

Including the interaction of lineup type and trial number did not further improve fit, relative to the model with just lineup type (see Table 4), $\chi^2(2) = 1.80$, $p = .41$, w_i ratio = 0.29, BF = 0.001; neither did including the interaction of disguise type and trial number, $\chi^2(3) = 3.40$, $p = .33$, w_i ratio = 0.22, BF = 3.20×10^{-5} ; nor did including the interaction of degree of disguise and trial number, $\chi^2(3) = 3.80$, $p = .28$, w_i ratio = 0.29, BF = 4.32×10^{-5} . Model fit was also not improved by allowing the slope of trial number to vary with memory strength, $\chi^2(4) = 1.80$, $p = .77$, w_i ratio = 0.04, BF = 2.91×10^{-7} , and note that this model failed to converge. Fit also did not improve by allowing the slope of trial number to vary by participants, $\chi^2(5) = 4.20$, $p = .52$, w_i ratio = 0.05, BF = 1.96×10^{-8} . Thus, confidence in correct identifications was best predicted by a three-level model that included the fixed effect of lineup type. Neither trial number nor any interactions with trial number predicted participants' confidence in correct identifications.

Confidence in correct rejections

The three-level null model was preferred over a one- or two-level null model. Confidence in correct rejections was highest when participants had a good memory strength ($M = 72.41\%$ [71.07, 73.75]), followed by a moderate memory strength ($M = 67.27\%$ [65.18, 69.36]), and a poor memory strength ($M = 63.36\%$ [61.50, 65.21]).

The fit was better than the null model when we included lineup type only, $\chi^2(1) = 10.60$, $p = .001$, w_i ratio = 77.48, BF = 4.26 (see Table 5). Confidence in correct rejections was higher when participants viewed simultaneous lineups ($M = 70.42\%$ [69.14, 71.70]), as compared to sequential lineups ($M = 65.68\%$ [64.17, 67.20]). Neither disguise type, $\chi^2(1) = 0.00$, $p = 1.00$, w_i ratio = 0.37, BF = 0.02; nor degree of disguise, $\chi^2(1) = 0.00$, $p = 1.00$, w_i ratio = 0.37, BF = 0.02; nor trial number (see Fig. 2), $\chi^2(1) = 0.00$, $p = 1.00$, w_i ratio = 0.37, BF = 0.02, improved the model fit.

Fit was not improved relative to the model with lineup type when we separately included the interactions of trial number with lineup type, $\chi^2(2) = 1.40$, $p = .50$, w_i ratio = 0.26, BF = 0.001; with disguise type, $\chi^2(3) = 0.60$, $p = .90$, w_i ratio = 0.07, BF = 1.01×10^{-5} ; and with degree of disguise, $\chi^2(3) = 0.00$, $p = 1.00$, w_i ratio = 0.05, BF = 7.89×10^{-6} . Fit was also not improved by allowing different slopes for each memory strength condition, $\chi^2(4) = 0.20$, $p = 1.00$, w_i ratio = 0.02, BF = 1.68×10^{-7} , or for each participant, $\chi^2(5) = 0.40$, $p = 1.00$, w_i ratio = 0.01, BF = 3.40×10^{-9} . In summary, the best-fitting model for predicting confidence in correct rejections was a three-level model with the fixed effect of lineup type. Neither trial number nor any interactions with trial number predicted participants' confidence in correct rejections.

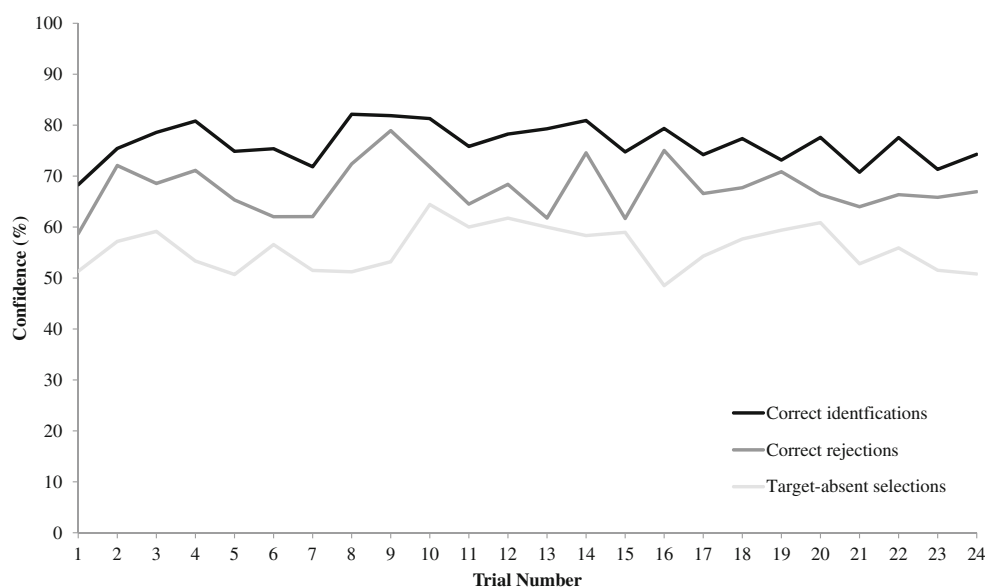


Fig. 2 Actual mean confidence (%) in correct identifications, correct rejections, and target-absent selections (i.e., any selection from a target-absent lineup) by trial number

Confidence in target-absent selections

A three-level model was appropriate for these data. We found that confidence in target-absent selections was highest when participants had a good memory strength ($M = 64.67\%$ [62.62, 66.72]), followed by a moderate memory strength ($M = 55.19\%$ [52.83, 57.55]), and a poor memory strength ($M = 53.26\%$ [51.80, 54.71]).

Lineup type alone, $\chi^2(1) = 16.00$, $p < .001$, w_i ratio = 1152.86, $BF = 77.48$, led to a better fit than the null model (see Table 6). The fit was not improved by adding disguise type, $\chi^2(1) = 1.40$, $p = .24$, w_i ratio = 0.78, $BF = 0.05$; degree of disguise, $\chi^2(1) = 2.20$, $p = .14$, w_i ratio = 1.22, $BF = 0.08$; or trial number (see Fig. 2), $\chi^2(1) = 2.60$, $p = .11$, w_i ratio = 1.49, $BF = 0.10$. Confidence in target-absent selections was higher when participants viewed sequential ($M = 59.91\%$ [58.27, 61.55]) rather than simultaneous ($M = 53.45\%$ [52.02, 54.89]) lineups.

The model with the interaction of lineup type and trial number was not significantly better than the model with lineup type alone, $\chi^2(2) = 2.80$, $p = .25$, w_i ratio = 0.58, $BF = 0.002$. The fit also did not improve when we included the interaction of trial number with disguise type, $\chi^2(3) = 4.20$, $p = .24$, w_i ratio = 0.39, $BF = 0.0001$, or the interaction of trial number with degree of disguise, $\chi^2(3) = 7.80$, $p = .0503$, w_i ratio = 2.46, $BF = 0.001$. Nor did the model fit improve through allowing the slope of trial number to vary with memory strength, $\chi^2(4) = 2.80$, $p = .59$, w_i ratio = 0.07, $BF = 1.37 \times 10^{-6}$, or by participant, $\chi^2(5) = 2.80$, $p = .73$, w_i ratio = 0.03, $BF = 3.39 \times 10^{-8}$. Thus, the best-fitting model for confidence in target-absent selections had three levels and included the fixed effect of lineup type. Neither trial number nor any

interactions with trial number predicted participants' confidence in target-absent selections.

Discussion

The purpose of this article was to determine whether it is appropriate to study lineup decisions and confidence in lineup decisions with a multiple-trial method. The typical eyewitness paradigm involving a single lineup decision and confidence rating per participant is resource-intensive and costly. However, using a multiple-trial approach would be ill advised if the effects of key variables of interest on eyewitness decisions are obscured in the data collected using this approach. Our results are good news for eyewitness researchers. That is, our results indicate that there is essentially no downside to using multiple lineup trials for different targets to examine accuracy, choosing, and confidence across manipulations of memory strength, disguise type, degree of disguise, and lineup type. Our most important results in this regard are that, with one exception, trial number did not interact with other variables of interest such as lineup type and memory strength. In the case of the exception, the effect size was negligible and therefore unlikely to influence the conclusions researchers draw using a multiple-trial approach. Main effect variations across trials should not be critical if other manipulated variables randomly vary or are counterbalanced across trials—but note that we found only one significant main effect and its effect size was trivial.

On the basis of the literature examining practice lineups (e.g., Shapiro & Penrod, 1986), we considered whether correct identifications would increase during the early trials but disappear in later trials, or whether there would be no effect

Table 4 Parameter estimates for predictors in models of confidence in correct identifications (2,683 observations).

Predictor	Model 1	Model 2 [†]	Model 3	Model 4	Model 5	Model 7	Model 8	Model 9	Model 10 [‡]	Model 11
Fixed Effects										
Intercept	76.84 (0.84)	74.67 (0.99) 4.62 (1.31)	77.53 (1.07)	77.76 (1.19)	77.68 (1.10)	75.48 (1.36) 4.73 (1.94)	76.56 (1.53) 4.67 (1.31) -2.12 (1.92)	77.17 (1.67) 2.94 (2.15)	75.53 (1.21) 4.63 (1.31)	75.54 (1.19) 4.62 (1.31)
Lineup type										
Disguise type			-1.36 (1.33)							
Degree of disguise				-0.37 (0.34)				-0.66 (0.46)		
Trial number					-0.07 (0.06)	-0.06 (0.08)	-0.09 (0.08)	-0.07 (0.06)	-0.07 (0.06)	-0.07 (0.06)
Lineup Type × Trial Number					-0.01 (0.11)		0.05 (0.11)			
Disguise Type × Trial Number										
Degree of Disguise × Trial Number								0.67 (0.68)		
Random Parameters										
Level 2 intercept variance (participant)	105.05 (10.25)	100.41 (10.02)	104.39 (10.22)	104.84 (10.24)	105.04 (10.25)	100.39 (10.02)	99.52 (9.98)	100.11 (10.00)	100.38 (10.02)	76.62 (8.75)
Level 2 slope variance (participant)										0.01 (0.10)
Level 3 intercept variance (memory strength)	0.74 (0.86)	0.50 (0.71)	0.74 (0.86)	0.75 (0.86)	0.74 (0.86)	0.50 (0.71)	0.50 (0.71)	0.51 (0.71)	0.22 (0.47)	0.56 (0.75)
Level 3 intercept variance (memory strength)*									0.27 (0.52)	
Level 3 slope variance (memory strength)									<0.001 (<0.001)	<0.001 (0.003)
Model Fit										
Model df	4	5	5	5	5	7	8	8	9	10
Test change in df	–	1 ^a	1 ^a	1 ^a	1 ^a	2 ^b	3 ^b	3 ^b	4 ^b	5 ^b
Bayes factor	–	8.58 ^a	0.03 ^a	0.04 ^a	0.04 ^a	0.001 ^b	3.20 × 10 ^{-5b}	4.32 × 10 ^{-5b}	2.91 × 10 ^{-7b}	1.96 × 10 ^{-8b}
AIC	23,765.1	23,754.9	23,766.1	23,765.9	23,765.7	23,757.4	23,757.9	23,757.4	23,761.4	23,761.0
Akaike weight	.003	.53	.002	.002	.002	.15	.12	.15	.02	.02
-2* log likelihood	23,757.1	23,744.9	23,756.1	23,755.9	23,755.7	23,743.4	23,741.9	23,741.4	23,743.4	23,741.0

Note: Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] indicates the best-fitting model.

* Model failed to converge; therefore the values should be interpreted with caution. Superscripts indicate df and Bayes factor for the comparison between the current model and ^a Model 1 (null model) and ^b Model 2. Model 6 is not included, since there was no need to test a model with multiple fixed effects; that is, only one fixed effect improved the model fit relative to the null model. *Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in Appendix A.

Table 5 Parameter estimates for predictors in models of confidence in correct rejections (2,531 observations)

Predictor	Model 1	Model 2 [†]	Model 3	Model 4	Model 5	Model 7	Model 8	Model 9	Model 10	Model 11
Fixed Effects										
Intercept	66.52 (2.47)	69.39 (2.63)	66.61 (2.63)	66.47 (2.66)	66.53 (2.60)	70.39 (2.90)	70.04 (3.00)	69.56 (3.12)	69.40 (2.59)	69.45 (2.58)
Lineup type		-5.69 (1.73)				-7.60 (2.40)	-7.04 (2.48)	-6.10 (2.66)	-5.68 (1.73)	-5.70 (1.73)
Disguise type			-0.17 (1.75)				-1.33 (2.45)			
Degree of disguise				0.02 (0.40)				-0.07 (0.58)		
Trial number					-0.001 (0.07)	-0.08 (0.10)	0.001 (0.07)	0.001 (0.07)	-0.002 (0.07)	-0.004 (0.07)
Lineup Type × Trial Number						0.15 (0.13)				
Disguise Type × Trial Number							2.61 (3.46)			
Degree of Disguise × Trial Number								0.16 (0.80)		
Random Parameters										
Level 2 intercept variance (participant)	187.96 (13.71)	180.49 (13.44)	187.92 (13.71)	187.95 (13.71)	187.95 (13.71)	181.07 (13.46)	180.22 (13.42)	180.49 (13.44)	180.35 (13.43)	173.22 (13.16)
Level 2 slope variance (participant)										
Level 3 intercept variance (memory strength)	15.83 (3.98)	16.17 (4.02)	15.83 (3.98)	15.83 (3.98)	15.83 (3.98)	16.32 (4.04)	15.88 (3.98)	16.18 (4.02)	13.51 (3.68)	13.42 (3.66)
Level 3 slope variance (memory strength)*									<0.001 (0.001)	<0.001 (0.002)
Level 3 slope variance (memory strength)									0.001 (0.03)	0.001 (0.03)
Model Fit										
Model df	4	5	5	5	5	7	8	8	9	10
Test change in df	-	1 ^a	1 ^a	1 ^a	1 ^a	2 ^b	3 ^b	2 ^b	4 ^b	5 ^b
Bayes Factor	-	4.26 ^a	0.02 ^a	0.02 ^a	0.02 ^a	0.001 ^b	1.01 × 10 ^{-5b}	7.89 × 10 ^{-6b}	1.68 × 10 ^{-7b}	3.40 × 10 ^{-9b}
AIC	23,256.9	23,248.2	23,258.9	23,258.9	23,258.9	23,250.9	23,253.6	23,254.1	23,256.0	23,257.9
Akaike weight	.01	.70	.003	.003	.003	.18	.05	.04	.01	.006
-2* log likelihood	23,248.9	23,238.2	23,248.9	23,248.9	23,248.9	23,236.9	23,237.6	23,238.1	23,238.0	23,237.9

Note: Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] Best-fitting model. Superscripts indicate df and Bayes factor for the comparison between the current model and ^a Model 1 (null model) and ^b Model 2. Model 6 is not included, since there was no need to test a model with multiple fixed effects; that is, only one fixed effect improved the model fit relative to the null model. *Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in [Appendix A](#).

Table 6 Parameter estimates for predictors in models of confidence in target-absent selections (1,657 observations)

Predictor	Model 1	Model 2 [†]	Model 3	Model 4	Model 5	Model 7	Model 8	Model 9	Model 10	Model 11
Fixed Effects										
Intercept	58.21 (2.61)	55.04 (2.70)	57.16 (2.75)	56.59 (2.84)	56.76 (2.76)	53.90 (2.92)	52.41 (3.08)	53.59 (3.16)	53.56 (2.75)	53.57 (2.76)
Lineup type		6.98 (1.72)				6.23 (2.47)	6.93 (1.72)	3.20 (2.76)	6.96 (1.72)	6.97 (1.72)
Disguise type			2.11 (1.75)				2.48 (2.43)			
Degree of disguise				0.65 (0.42)				0.03 (0.54)		
Trial number					0.11 (0.07)	0.09 (0.09)	0.13 (0.10)	0.11 (0.07)	0.12 (0.07)	0.12 (0.07)
Lineup Type × Trial Number						0.06 (0.14)				
Disguise Type × Trial Number										
Degree of Disguise × Trial Number							−0.04 (0.14)	1.49 (0.86)		
Random Parameters										
Level 2 intercept variance (participant)	167.10 (12.93)	156.15 (12.50)	166.48 (12.90)	166.91 (12.92)	168.36 (12.98)	157.30 (12.54)	156.82 (12.52)	157.18 (12.54)	157.46 (12.55)	162.54 (12.75)
Level 2 slope variance (participant)										
Level 3 intercept variance (memory strength)	17.82 (4.22)	17.63 (4.20)	17.82 (4.22)	18.34 (4.28)	17.98 (4.24)	17.72 (4.21)	17.81 (4.22)	18.19 (4.26)	16.24 (4.03)	16.23 (4.03)
Level 3 intercept variance (memory strength)*									<0.001 (0.001)	
Level 3 slope variance (memory strength)									<0.001 (0.02)	<0.001 (0.02)
Model Fit										
Model df	4	5	5	5	5	7	8	8	9	10
Test change in df	–	1 ^a	1 ^a	1 ^a	1 ^a	2 ^b	3 ^b	3 ^b	4 ^b	5 ^b
Bayes factor	–	77.48 ^a	0.05 ^a	0.08 ^a	0.10 ^a	0.002 ^b	0.0001 ^b	0.001 ^b	1.37 × 10 ^{−6b}	3.39 × 10 ^{−8b}
AIC	14,703.1	14,689.0	14,703.6	14,702.7	14,702.3	14,690.1	14,690.9	14,687.2	14,694.3	14,696.2
Akaike weight	.0002	.22	.0002	.0002	.0003	.13	.09	.54	.02	.01
−2* log likelihood	14,695.1	14,679.0	14,693.6	14,692.7	14,692.3	14,676.1	14,674.9	14,671.2	14,676.3	14,676.2

Standard errors for fixed effects and standard deviations for random effects are given in parentheses. df = degrees of freedom. AIC = Akaike information criterion. [†] Best-fitting model. Superscripts indicate df and Bayes factor for the comparison between the current model and ^a Model 1 (null model) and ^b Model 2. Model 6 is not included, since there was no need to test a model with multiple fixed effects; that is, only one fixed effect improved the model fit relative to the null model. ^{*}Two intercepts are required to represent the interaction of memory strength and trial number because there were three levels of memory strength in our data (good, moderate, and poor). Model equations are available in [Appendix A](#).

overall. We did find a statistically significant quadratic effect of trial number on correct identifications but the size of the effect was negligible. That is, our results indicate that, although trial number was a significant predictor, a correct identification was essentially as likely on one trial as another ($OR = 0.99$). Likewise, the interaction of the quadratic effect of trial number and disguise type was a significant predictor of correct identifications, but again the effect size was negligible ($OR = 1.00$). Visual inspection of the main effect and interaction (See Fig. 1B) illustrates that the nature of these effects are almost impossible to discern. Given our highly powerful data set—our analysis of correct identifications included 4,188 data points—and the small effect sizes, we feel confident in concluding that multiple trials will not obscure the effects of other variables on correct identification rates. Thus, although we found practice effects on correct identifications, the effects are small and are highly unlikely to influence the conclusions researchers draw from multiple-trial experiments.

Contrary to our expectations, lineup type did not interact with trial number to influence correct rejections. Thus, researchers wanting to use a multiple-trial design with sequential lineups may do so despite the fact that participants may discern the size of the lineup. One potential explanation for the lack of the predicted effect may be that we terminated the sequential lineups whenever a selection was made. As a result, participants may not have deduced the number of lineup members in the sequential conditions because the apparent lineup size varied with their choosing behavior. We did not draw participants' attention to the number of lineup members in our sequential lineups and participants completed a randomized assortment of target-present and target-absent trials. An effect of lineup size may emerge if lineup size is more salient (e.g., Horry et al., 2012; Lindsay et al., 1991), but this could be countered by varying lineup size across trials.

Although it was beyond the scope of our research question, we found an interesting result with correct rejections from sequential and simultaneous lineups across our memory strength conditions. Correct rejections declined somewhat for sequential lineups from the good to poor memory strength conditions (.78, .69, .56, respectively); however, the decline for simultaneous lineups was much more dramatic (.70, .63, .41, respectively). We did not test for an interaction of memory strength and lineup type—and chose not to because participants were not randomly assigned to memory strength conditions—thus, we can only raise this as a possible avenue for future research. Evidence regarding how memory strength influences sequential and simultaneous lineup performance is likely to be highly relevant to the current debate over which procedure is superior (Wells, Smalarz, & Smith, 2015; Wells, Smith, & Smalarz, 2015; Wixted & Mickes, 2015a, 2015b).

We found no indication that participants with a good (versus poor) memory trace were more willing to identify someone as trials progressed. Trial number was not included in the

best-fitting model of choosing either as a fixed effect or in interaction with any of our manipulations related to memory trace (i.e., memory strength, disguise type, or degree of disguise). These data are consistent with our expectation that exposing participants to a random ordering of target-present and -absent trials limits the opportunity to inflate their perceived ability to identify targets. This finding further supports the use of multiple-trial experiments for studying eyewitness identification.

We found similarly encouraging results with confidence as with accuracy and choosing. Neither trial number, nor any interactions with trial number significantly predicted confidence ratings—regardless of decision type (i.e., correct identification, correct rejection, or selection from a target-absent lineup). The alternative, more conservative, fit indices (Akaike weights, BF) are consistent with the likelihood ratio test, on which we based our model selections, with one exception. The exception was confidence in target-absent selections in which the highest Akaike weight (and therefore the most likely best-fitting model, according to this approach; Wagenmakers & Farrell, 2004), was the three-level model with lineup type, degree of disguise, trial number, and the interaction of degree of disguise and trial number as predictors. Supplemental Fig. 1 provides a visual representation of the interaction of degree of disguise and trial number for confidence in target-absent selections. No systematic effect of trial number is readily apparent across or within memory strength conditions.

Despite the high level of power we had in this experiment—with 349 participants and within-subjects manipulations of degree of disguise and trial number resulting in 1,657 to 8,376 data points, depending on the dependent variable—we detected three possible effects of trial number. The first two effects were on correct identifications: a significant direct quadratic effect of trial number, with a negligible effect size, and an interaction of the quadratic effect of trial number with disguise type, also with a negligible effect size. We encourage readers to carefully consider the practical relevance of small but significant effects detected by using a liberal test given to a large sample. The second effect, on confidence in target-absent selections, emerged from only on one of the three model-fitting criteria we reported—Akaike weights—with no easily discernible systematic pattern. Thus, the preponderance of evidence suggests that multiple-trial experiments are appropriate for eyewitness identification experiments.

This study has two potential limitations. First, our manipulation of memory strength was confounded with data collection location and dates. That is, we collected the good and moderate memory strength data in Eastern Canada between 2005 and 2007, and the poor memory strength data in Western Canada between 2009 and 2010. Despite this, there is no logical reason to expect that location or date would systematically affect either lineup decisions or confidence. Indeed, the results with regard to memory strength are in line with traditional expectations (e.g., correct identifications, correct

rejections, and confidence in correct identifications were higher when memory strength was better); therefore, we think that the variability in our sample probably enhances our generalizability more than this confound negates it. Importantly, the other independent variables produced similar patterns of results in the three data sets, suggesting that participants behaved alike regardless of data collection date or location.

Second, this research examined only one type of multiple-trial identification design—presenting alternating mock-crime videos followed by the yoked lineups (i.e., Crime-Lineup-Crime-Lineup). Meissner and colleagues (Evans, Marcon, & Meissner, 2009; Lane & Meissner, 2008; Meissner et al., 2005) have used a different multiple-trial method in which participants view all targets before completing the accompanying lineups (i.e., Crime-Crime-Lineup-Lineup). It remains to be seen whether Meissner’s paradigm is comparable to our multiple-trial experiments and the standard, single-trial method. Indeed, basic memory research indicates that interference may build up across to-be-remembered lists, which is alleviated by testing (i.e., list-before-last paradigm; Jang & Huber, 2008; Klein, Shiffrin, & Criss, 2007; Shiffrin, 1970). Presenting multiple to-be-remembered target faces or mock-crimes before presenting any lineups may lead to a buildup of interference. In our paradigm, participants had to maintain a memory for only one target at a time, which may have prevented interference. Further research will be necessary before results from our paradigm can be generalized to Meissner and colleagues’ multiple-trial method.

In conclusion, researchers should consider using our multiple-trial paradigm (Crime-Lineup-Crime-Lineup) to obtain more data from fewer participants. Future research into multiple-trial designs should examine whether any systematic effects emerge beyond 24 trials or with any other paradigms (e.g., that of Meissner et al., 2005). Overall, a multiple-trial design for independent lineup trials can be an effective way of obtaining powerful datasets in lineup experiments, allowing researchers to examine more complex interactions than are typically tested and that could significantly contribute to our understanding of eyewitness decision making.

Appendix A

We used a logit link function for calculating the accuracy and choosing models because accuracy and choosing are dichotomous measures and the standard linear link function for calculating the confidence models because confidence is a continuous measure (Heck, Thomas, & Tabata, 2012). We used the lme4 package available for R for all multilevel modeling (Bates, Maechler, Bolker, & Walker, 2013). All multilevel models used maximum likelihood estimation, which is necessary for conducting likelihood

ratio tests, and the variance structure was unspecified. The default optimizer for lmer(), bobyqa, was used unless the model did not converge. In this situation, we attempted to fit the model using the optimizers that comprise the allFit() package for lme4 (Nelder_Mead, nlminbw, nmkbw, L-BFGS-B, NLOPT_LN_NELDERMEAD, NLOPT_LN_BOBYQA).

In these formulas, η_{tij} refers to the predicted outcomes for a measure (e.g., correct identifications) on trial t , for participant i , who was in memory strength condition j . We use β_{0ij} to represent the variation in random intercepts at Level 3 (memory strength), γ_{n00} to represent fixed-effect coefficients at Level 1 (trial level) where γ_{000} represents the Level 1 intercept and $\gamma_{100}, \gamma_{200}$, and so forth, refer to fixed-effect coefficients, μ_{0ij} to represent random effects at Level 2 (participant level), and μ_{00j} to represent random effects at Level 3 (memory strength level).

We began each analysis by checking that a multilevel model was appropriate. First we produced a model with no predictors other than an intercept at Level 3 (Eq. 1). Therefore,

$$\eta_{tij} = \gamma_{000} \quad (1)$$

We compared the initial null model to one with the fixed intercept at Level 3 and a random intercept at Level 2 (Eq. 2) and then to one with the fixed intercept at Level 3 and random intercepts at Levels 1 and 2 (Eq. 3). The best-fitting model amongst these three models was used as the null model for the subsequent modeling procedure. We thus followed the steps below to model our dependent measures.

$$\eta_{tij} = \gamma_{000} + u_{0ij} \quad (2)$$

$$\eta_{tij} = \gamma_{000} + u_{0ij} + u_{00j} \quad (3)$$

We used the following process to determine the best-fit model for each of our six measures:

1. We compared the null model to a fixed-effect model with a single predictor (X ; Eq. 4). The predictors used were lineup type (Model 2), disguise type (Model 3), degree of disguise (Model 4), and trial number (Model 5) as described in the methods section.

$$\eta_{tij} = \gamma_{000} + \gamma_{100}X_{tij} + u_{0ij} + u_{00j} \quad (4)$$

2. We compared the fit of a model containing all of the fixed effects that led to better fit than the null model to each of the fixed-effect models with one predictor that led to a better fit than the null model and the null model (Model 6; Eq. 5). The model from this analysis that fit the data best was named the “best-fitting fixed-effects model” and was used in subsequent steps.

$$\eta_{tij} = \gamma_{000} + \gamma_{100}X_{tij} + \dots + \gamma_{n00}X_{nij} + u_{0ij} + u_{00j} \quad (5)$$

3. We compared the best-fitting fixed-effects model to a model containing the same predictors as the best-fitting fixed-effects model plus the fixed interaction of lineup type and trial number (Model 7; Eq. 6).

$$\eta_{ij} = \gamma_{000} + \gamma_{100}\text{TrialNumber}_{ij} + \gamma_{200}\text{LineupType}_{ij} \\ + \gamma_{300}\text{TrialNumber} : \text{LineupType}_{ij} + \dots \gamma_{n00}X_{ij} \quad (6) \\ + u_{0ij} + u_{00j}$$

4. We compared the best-fitting fixed-effects model to a model with the same predictors as the best-fitting fixed-effects model plus the fixed interaction of disguise type and trial number (same as Step 3, but using disguise type in place of lineup type; Model 8; Eq. 7).

$$\eta_{ij} = \gamma_{000} + \gamma_{100}\text{TrialNumber}_{ij} + \gamma_{200}\text{DisguiseType}_{ij} \\ + \gamma_{300}\text{TrialNumber} : \text{DisguiseType}_{ij} + \dots \gamma_{n00}X_{ij} \quad (7) \\ + u_{0ij} + u_{00j}$$

5. We compared the best-fitting fixed-effects model to a model containing the same predictors as the best-fitting fixed-effects model plus the fixed interaction of degree of disguise and trial number (same as Steps 3 and 4, but using degree of disguise instead of lineup type or disguise type; Model 9; Eq. 8).

$$\eta_{ij} = \gamma_{000} + \gamma_{100}\text{TrialNumber}_{ij} + \gamma_{200}\text{DisguiseLevel}_{ij} \\ + \gamma_{300}\text{TrialNumber} : \text{DisguiseLevel}_{ij} + \dots \gamma_{n00}X_{ij} \quad (8) \\ + u_{0ij} + u_{00j}$$

6. If model fit had improved by adding more than one interaction (Steps 3–5) we would have compared the individual interaction models to a model containing all of the interactions; however, this never occurred in our data. In the next step, we compared the best-fitting fixed-effects model (from Steps 1–5) with a model containing the same predictors as the best-fitting fixed-effects model but that also allowed trial number to vary with memory strength condition (interaction of trial number and memory strength). This model (Eq. 9) is the same as that above (Eq. 8), but includes a random effect for trial number at Level 3 (u_{10j} ; Model 10).

$$\eta_{ij} = \gamma_{000} + \gamma_{100}\text{trialNumber}_{ij} + \dots \gamma_{n00}X_{ij} + u_{00j} \\ + u_{0ij} + u_{10j}\text{trialNumber}_{10j} \quad (9)$$

7. We compared the best-fitting model (from Steps 1–6) with a model containing the same predictors as the best-fitting model and that allowed trial number to vary for each participant (interaction of trial number and participant). This model (Eq. 10) is the

same as Eq. 9 with a random effect for trial number at Level 2 (u_{1ij} ; Model 11).

$$\eta_{ij} = \gamma_{000} + \gamma_{100}\text{trialNumber}_{ij} + \dots \gamma_{n00}X_{ij} + u_{00j} \\ + u_{0ij} + u_{10j}\text{trialNumber}_{10j} + u_{1ij}\text{trialNumber}_{1ij} \quad (10)$$

Appendix B: Flow chart for our model-testing procedure

*Note: Step numbers correspond with steps/equations described in Appendix A

Determine whether multilevel modeling is appropriate

- Does a **two-level** null model fit the data better than a **one-level** null model?
 - o Yes – Does a **three-level** null model fit the data better than a **two-level** null model?
 - Yes – Retain the three-level null model as the null model in the following steps (Model 1)
 - No – Retain the three-level null model as the null model in the following steps (Model 1)
 - o No – Conduct a regression analysis

1. Determine which predictors improve fit relative to the null model

- Does a model with **lineup type** only fit the data better than the null model? (Model 2)
 - o Yes – Include lineup type in next step
 - o No – Do not include lineup type in next step
- Does a model with **disguise type** only fit the data better than the null model? (Model 3)
 - o Yes – Include disguise type in next step
 - o No – Do not include disguise type in next step
- Does a model with **degree of disguise** only fit the data better than the null model? (Model 4)
 - o Yes – Include degree of disguise in next step
 - o No – Do not include degree of disguise in next step
- Does a model with **trial number** only fit the data better than the null model? (Model 5)
 - o Yes – Include trial number in next step
 - o No – Do not include trial number in next step

Determine whether a model with multiple fixed-effect predictors is better than a model with one fixed-effect predictor

2. Does a model with “yes” predictors from the previous step perform better than a model with a single predictor? (Model 6)
 - o Yes – Incorporate all predictors into the best-fitting fixed-effects model

- o No – Compare the single predictor models and retain the best-fitting as the best-fitting fixed-effects model

Determine whether model fit can be improved by adding fixed effect interactions at Level 1

3. Is fit of the best-fitting fixed-effects model improved by adding the fixed effect **interaction of lineup type and trial number**? (Model 7)
 - a Yes – Retain interaction for next step
 - b No – Do not retain interaction for next step
4. Is fit of the best-fitting fixed-effect model improved by adding the fixed effect **interaction of disguise type and trial number**? (Model 8)
 - a Yes – Retain the interaction for the next step
 - b No – Do not retain the interaction for the next step
5. Is fit of the best-fitting fixed-effect model improved by adding the fixed effect **interaction of degree of disguise and trial number**? (Model 9)
 - a Yes – Retain the interaction for the next step
 - b No – Do not retain the interaction for the next step

Determine whether a model with multiple interactions is better than a model with one interaction

(Note: this step never occurred because we never found more than one interaction that improved model fit)

- Does a model with “yes” interactions from the previous step perform better than a model with a single interaction?
- Yes – Incorporate all interactions into the best-fitting fixed-effects model
 - No – Compare interaction models and retain the best-fitting as the best-fitting fixed-effects model

Determine whether model fit can be improved by allowing the slope of trial number to vary

6. Does model fit improve with the addition of the **interaction of trial number and memory strength**? (Model 10)
 - a Yes – Incorporate the interaction into the best-fitting model
 - b No – Retain the previous best-fitting model
7. Does model fit improve with the addition of the **interaction of trial number and participant**? (Model 11)
 - a Yes – Incorporate the interaction into the best-fitting model
 - b No – Retain the previous best-fitting model

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.0-5) [Computer Software]. Retrieved from <http://cran.r-project.org/package=lme4>
- Brewer, N., Weber, N., Clark, A., & Wells, G. L. (2008). Distinguishing accurate from inaccurate eyewitness identifications with an optional deadline procedure. *Psychology, Crime & Law*, 14, 397–414. doi:10.1080/10683160701770229
- Council, N. R. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: National Academies Press.
- Evans, J. R., Marcon, J. L., & Meissner, C. A. (2009). Cross-racial lineup identification: Assessing the potential benefits of context reinstatement. *Psychology Crime and Law*, 15, 19–28. doi:10.1080/10683160802047030
- Feather, N. T. (1966). Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology*, 3, 287–298. doi:10.1037/h0022965
- Field, A. (2009). Multilevel linear models. In *Discovering statistics using SPSS (and sex, drugs, and rock 'n' roll)* (pp. 725–778). London, UK: Sage.
- Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the life span: A meta-analysis of age differences. *Psychological Bulletin*, 141, 1228–1265. doi:10.1037/bul0000013
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
- Godfrey, R. D., & Clark, S. E. (2010). Repeated eyewitness identification procedures: Memory, decision making, and probative value. *Law and Human Behavior*, 34, 241–258. doi:10.1007/s10979-009-9187-7
- Goodman, G. S., Bottoms, B. L., Schwartz-Kenney, B. M., & Rudy, L. (1991). Children's testimony about a stressful event: Improving children's reports. *Journal of Narrative and Life History*, 1, 69–99.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2012). *Multilevel modeling of categorical outcomes using IBM SPSS*. New York, NY: Routledge.
- Horry, R., Halford, P., Brewer, N., Milne, R., & Bull, R. (2014). Archival analyses of eyewitness identification test outcomes: What can they tell us about eyewitness memory? *Law and Human Behavior*, 38, 94–108. doi:10.1037/lhb0000060
- Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied*, 18, 346–360. doi:10.1037/a0029779
- Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Innocence Project. (n.d.). Retrieved September 26, 2016, from www.innocenceproject.org
- Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 112. doi:10.1037/0278-7393.34.1.112
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes Factors. *Journal of Problem Solving*, 7, 2–9. doi:10.7771/1932-6246.1167
- Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. Nairne (Ed.), *The Foundations of Remembering: Essays in honor of Henry L. Roediger, III* (pp. 171–189). New York, NY: Psychology Press.
- Lane, S. M., & Meissner, C. A. (2008). A “middle road” approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology*, 22, 779–787. doi:10.1002/acp.1482

- Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (pp. 377–425). Mahwah, NJ: Erlbaum.
- Lindsay, R. C. L., Lea, J. A., & Fulford, G. J. (1991). Sequential lineup presentation: Technique matters. *Journal of Applied Psychology*, 76, 741–745. doi:10.1037/0021-9010.76.5.741
- Lindsay, R. C. L., Mansour, J. K., Beaudry, J. L., Leach, A.-M., & Bertrand, M. I. (2009). Sequential lineup presentation: Patterns and policy. *Legal & Criminological Psychology*, 14, 13–24. doi:10.1348/135532508X382708
- Lindsay, R. C. L., Mansour, J. K., Kalmet, N., Bertrand, M. I., & Melsom, E. (2011). Face perception and recognition in eyewitness memory. In A. Calder, G. Rhodes, M. Johnson, J. Haxby, & J. Keane (Eds.), *The handbook of face perception* (pp. 307–328). Oxford, UK: Oxford University Press.
- Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, 9, 215–218. doi:10.1111/1467-9280.00041
- Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70, 556–564. doi:10.1037/0021-9010.70.3.556
- Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, 66, 482–489. doi:10.1037/0021-9010.66.4.482
- Malpass, R. S., Laviqueur, H., & Weldon, D. E. (1973). Verbal and visual training in face recognition. *Attention, Perception, and Psychophysics*, 14, 285–292. doi:10.3758/BF03212392
- Mansour, J. K., Beaudry, J. L., Bertrand, M. I., Kalmet, N., Melsom, E., & Lindsay, R. C. L. (2012). Impact of disguise on identification decision and confidence with simultaneous and sequential lineups. *Law and Human Behavior*, 36, 513–526. doi:10.1037/h0093937
- Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behavior on a lineup task. *Applied Cognitive Psychology*, 23, 1012–1026. doi:10.1002/acp.1570
- Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition*, 33, 783–792. doi:10.3758/BF03193074
- Palmer, M. A., Brewer, N., & Weber, N. (2010). Post-identification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied*, 16, 387–398.
- Parker, J. F., & Myers, A. (2001). Attempts to improve children's identifications from sequential presentation lineups. *Journal of Applied Social Psychology*, 21, 796–815. doi:10.1111/j.1559-1816.2001.tb01414.x
- Parker, J. F., & Ryan, V. (1993). An attempt to reduce guessing behavior in children's and adult's eyewitness identifications. *Law and Human Behavior*, 17, 11–26. doi:10.1007/BF01044534
- Platz, S. J., & Hosch, H. M. (1988). Cross-racial/ethnic eyewitness identification: A field study. *Journal of Applied Social Psychology*, 18, 972–984. doi:10.1111/j.1559-1816.1988.tb01187
- Pozzulo, J. D., & Lindsay, R. C. L. (1997). Increasing correct identifications by children. *Expert Evidence*, 5, 126–132. doi:10.1023/A:1008875802767
- Pozzulo, J. D., & Lindsay, R. C. L. (1998). Identification accuracy of children versus adults: A meta-analysis. *Law and Human Behavior*, 22, 549–570. doi:10.1023/A:1025739514042
- Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.
- Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, 100, 139–156. doi:10.1037/0033-2909.100.2.139
- Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science*, 168, 1601–1603. doi:10.1126/science.168.3939.1601
- State v. Henderson*, 208 N.J. 208 (2011).
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Montreal, PQ: Allyn & Bacon.
- Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement* (Office of Justice Programs Document No. NCJ-178240). Washington, DC: United States Department of Justice.
- Turtle, J. W., Lindsay, R. C. L., & Wells, G. L. (2003). Best practice recommendations for eyewitness evidence procedures: New ideas for the oldest way to solve a case. *Canadian Journal of Police and Security Services*, 1, 5–18.
- VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, 47, 513–539. doi:10.1146/annurev.psych.47.1.513
- Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11, 192–196. doi:10.3758/BF03206482
- Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27, 359–397. doi:10.1177/0049124199027003002
- Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, 4, 313–317. doi:10.1016/j.jarmac.2015.08.008
- Wells, G. L., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition*, 4, 324–328. doi:10.1016/j.jarmac.2015.08.010
- Wixted, J. T., & Mickes, L. (2015a). Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition*, 4, 318–323. doi:10.1016/j.jarmac.2015.08.009
- Wixted, J. T., & Mickes, L. (2015b). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, 4, 329–334. doi:10.1016/j.jarmac.2015.08.007
- Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, 62, 439–456.